Information, data dimension and factor structure

Jan P.A.M. Jacobs*
University of Groningen

Pieter W. Otter
University of Groningen

Ard H.J. den Reijer
Sveriges Riksbank

November 28, 2011

Abstract

This paper employs concepts from information theory to choosing the dimension of a data set. We propose a relative information measure connected to Kullback-Leibler numbers. By ordering the series of the data set according to the measure, we are able to obtain a subset of a data set that is most informative. The method can be used as a first step in the construction of a dynamic factor model or a leading index, as illustrated with a Monte Carlo study and with the U.S. macroeconomic data set of Stock and Watson [21].

Keywords: Kullback-Leibler numbers, information, factor structure, data set dimension, dynamic factor models, leading index

JEL-code: C32, C52, C82

*Corresponding author: Jan P.A.M. Jacobs, Faculty of Economics and Business, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands. Tel.: +31 50 363 3681. Fax: +31 50 363 7337. Email: `j.p.a.m.jacobs@rug.nl`

1 Introduction

With the proliferation of huge data sets a natural question to ask is how much information there is in a data set. Is there an 'optimal' size of the data set in relation to some variable(s) of interest, in other words can we confine attention to a subset of the series instead of having to monitor all series in a data set? The question seems especially relevant for factor models, which exploit the idea that movements in a large number of series are driven by a limited number of common 'factors'. For a recent overview see Bai and Ng [4].

Although convergence of factor estimates requires large cross-sections and large time dimensions, see e.g. Forni and Lippi [9] and Bai [1], the data set need not be very large to obtain reasonably precise factor estimates. Boivin and Ng [6] and Inklaar, Jacobs, and Romp [13] find that some 40 variables are sufficient using Monte Carlo simulations and a comparison to conventional NBER-type business cycle indicators, respectively. Bai and Ng [2] also conclude that the number of series need not be very large to get precise factor estimates. The question whether we can confine attention to a subset of the variables is also relevant for the construction of leading indexes, which aims at selecting indicators with predictive power out of a large number of candidates too.[1]

Building upon Otter and Jacobs [18], the paper exploits concepts from information theory, in particular Kullback-Leibler numbers, to analyse infor-

[1] Another issue in the construction of (dynamic) factor models is the determination of the number of factors. For a discussion of the literature and a criterion for the determination of the number of factors see Otter, Jacobs and den Reijer [19].

mation in the data.[2] We propose a relative information measure based on Gaussian distributed data with a clear link to Kullback-Leibler numbers. The measure is discussed in more detail assuming an approximate factor structure in the data. A recursive procedure including a test is given whether an additional variable adds information. Ordering the series of the data set according to the measure enables us to identify a subset of a data set that is most informative. The method can be used as a first step in the construction of a dynamic factor model or a leading index.

Our paper is related to Bai and Ng [5], who study 'hard' and 'soft' thresholding to reduce the influence of uninformative predictors for a variable from the point of view of factor forecasting. Hard thresholding involves some pretest procedure, while under soft thresholding the top ranked predictors according to some soft-thresholding rule are kept. Our paper fits into the category of soft thresholding; we also seek to identify a subset of a larger data set that is most informative. However, in contrast with the penalized regression models studied by Bai and Ng [5], the Least Absolute Shrinkage Selection Operator (LASSO) model of Tibshirani [22] and the elastic net rule of Zou and Hastie [24], our method is based on a quantitative measure of information adopting a factor model framework and does not rely on an external regression method.

We illustrate the concepts with a Monte Carlo simulation and with the macroeconomic data set of Stock and Watson [21], which consists of 132 monthly U.S. variables and runs from 1959–2003. We find that relative in-

[2] Jacobs and Otter [14] apply similar information concepts to derive a formal test for the number of common factors and the lag order in a dynamic factor model.

formation is indeed maximized for a limited number of series. In the Stock and Watson data set relative information is maximized for 40–50 series, if we are interested in modelling industrial production and CPI inflation.

The paper is structured as follows. Section 2 discusses our relative information measure, how it works out assuming an approximate factor structure in the data, and presents a test procedure. After a Monte Carlo study in Section 3, we apply our method to the U.S. data set of Stock and Watson [21] in Section 4. Section 5 concludes.

2 Information in data

2.1 Kullback-Leibler numbers and information

Let $f_1(\tilde{\boldsymbol{x}}) : \tilde{\boldsymbol{x}} \sim \mathcal{N}_N(\mathbf{0}, \boldsymbol{\Gamma} = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}')$ be the density function of an N-dimensional data vector $\boldsymbol{x}$ (time index suppressed), then $f_1(\boldsymbol{x}) : \boldsymbol{x} \sim \mathcal{N}_N(\mathbf{0}, \boldsymbol{\Lambda})$ where $\boldsymbol{x} = \boldsymbol{C}'\tilde{\boldsymbol{x}}$. Let $f_2(\tilde{\boldsymbol{x}}) : \tilde{\boldsymbol{x}} \sim \mathcal{N}_N(\mathbf{0}, \boldsymbol{I}_N)$. Then $f_2(\boldsymbol{x}) : \boldsymbol{x} \sim \mathcal{N}_N(\mathbf{0}, \boldsymbol{I}_N)$ with $\boldsymbol{x} = \boldsymbol{C}'\tilde{\boldsymbol{x}}$. The so-called *Kullback-Leibler* numbers are defined as

$$G_1 = \mathrm{E}_{f_1}\left(\log\left(\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})}\right)\right) \text{ and } G_2 = \mathrm{E}_{f_2}\left(\log\left(\frac{f_2(\boldsymbol{x})}{f_1(\boldsymbol{x})}\right)\right), \tag{1}$$

and $G = G_1 + G_2$ is the measure of information for discriminating between the two density functions with $G = 0$ in case $f_1(\boldsymbol{x}) = f_2(\boldsymbol{x})$ and $G = \infty$ in case of perfect discrimination, see Young and Calver [23], p245. For a general background see Burnham and Anderson [7].

For $\mathrm{tr}(\boldsymbol{\Gamma}) = \mathrm{tr}(\boldsymbol{\Lambda}) = N$ we have $G_1 = -\mathrm{logdet}(\boldsymbol{\Lambda})$, where G_1 is the mean information in $\boldsymbol{x}$ for discriminating between $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$, see Kullback and Leibler [15], and $G_2 = \text{logdet}(\boldsymbol{\Lambda}) + \frac{1}{2}\left(\text{tr}(\boldsymbol{\Lambda}^{-1}) - N\right)$. Therefore

$$2G = \text{tr}(\boldsymbol{\Lambda}^{-1}) - N = \text{tr}(\boldsymbol{\Lambda}^{-1}) - \text{tr}(\boldsymbol{\Lambda}) = \sum_{j=1}^{N} \frac{(1-\lambda_j^2)}{\lambda_j} = \sum_{j=1}^{N} \frac{(1-\lambda_j)(1+\lambda_j)}{\lambda_j}, \tag{2}$$

from which it can be seen that G is small (not discriminating) if the eigenvalues λ_j are close to 1, but becomes large (discriminating) for "small" eigenvalues.

We can also use the entropy measure. Let $\boldsymbol{x}_t$ again be an N-dimensional vector of observed data at time t, $t = 1, \ldots, T$. The data is demeaned and normalized, and normally distributed with mean zero and variance $\text{E}(\boldsymbol{x}_t \boldsymbol{x}_t') = \boldsymbol{\Gamma}$, i.e. $\boldsymbol{x}_t \sim \mathbb{N}(\boldsymbol{0}, \boldsymbol{\Gamma})$, where $\text{diag}(\boldsymbol{\Gamma}) = (1, 1, \ldots, 1)$ and $\text{tr}(\boldsymbol{\Gamma}) = N$. Here we make the additional assumption that all eigenvalues are positive. The entropy as measure of disorder for a stationary, normally distributed vector is given by

$$2H_x = -2\,\text{E}_x\left[\log f(\boldsymbol{x})\right] = cN + \text{logdet}(\boldsymbol{\Gamma}),$$

where $c \equiv \log(2\pi) + 1 \approx 2.84$, with $2H_{x,max} = cN$ in case $\boldsymbol{\Gamma} = \boldsymbol{I}_N$, see e.g. Goodwin and Payne (1977) [10]. The information or negentropy is defined as

$$2\text{Inf}_x \equiv 2(H_{x,max} - H_x) = -\text{logdet}(\boldsymbol{\Gamma}) \geq 0, \tag{3}$$

which is zero in case $\boldsymbol{\Gamma} = \boldsymbol{I}_N$. This measure coincides with Kullback-Leibler information G_1. We define the *relative information* as

$$\text{Inf}_N^R = \frac{2H_{max} - 2H_{x(N)}}{2H_{max}} = \frac{2\text{Inf}_N}{2H_{max}} = \frac{2\text{Inf}_N}{cN}. \tag{4}$$

If $H_{x(N)}$ is equal to H_{max} then $\mathrm{Inf}_N^R = 0$; if $H_{x(N)} = 0$ then $\mathrm{Inf}_N^R = 1$. The relative information equals the weighted mean information per variable in the data vector $\boldsymbol{x}_t$, where the weight is $1/c$.

2.2 Relative information measure Inf_n^R in the approximate factor model

In this section we consider the relative information measure in more detail assuming an approximate factor structure in the data. Let the n-dimensional data vector $\boldsymbol{x}_t$ be driven by k factors

$$\boldsymbol{x}_t = \boldsymbol{B}_n \boldsymbol{F}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{x}_t \in \mathbb{R}^n, \ \boldsymbol{F}_t \sim \mathcal{N}_k\left(\boldsymbol{0}, \boldsymbol{I}_k\right), \ \boldsymbol{\varepsilon}_t \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{\Psi}_{11}), \tag{5}$$

where $\boldsymbol{B}_n \in \mathbb{R}^{n \times k}$ is the matrix of factor loadings, and the idiosyncratic errors $\boldsymbol{\varepsilon}_t$ are allowed to be 'weakly' correlated across n and t. Since a dynamic factor model with q factors and p lags can be written as a static factor models with $r = q(p+1)$ factors (see e.g. Bai and Ng [4], Section 2), the approximate factor model of Equation (5) is sufficiently general to cover the static and the dynamic case. The generalized dynamic factor structure of Forni and Lippi [9] and Forni et al. [8] can be dealt with too.

The variance between the n elements of $\boldsymbol{x}_t$ is equal to $\boldsymbol{\Gamma}(n) = \boldsymbol{B}_n \boldsymbol{B}_n' + \boldsymbol{\Psi}_{11}$. Adding a variable $x_{n+1,t}$ we have

$$\begin{pmatrix} \boldsymbol{x}_t \\ x_{n+1,t} \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}_n \\ \boldsymbol{b}_{n+1} \end{pmatrix} \boldsymbol{F}_t + \begin{pmatrix} \boldsymbol{\varepsilon}_t \\ \varepsilon_{n+1,t} \end{pmatrix}, \tag{6}$$

with covariance $\boldsymbol{\Gamma}(n+1) = \begin{pmatrix} \boldsymbol{\Gamma}(n) & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & 1 \end{pmatrix}$, where $\boldsymbol{\Gamma}_{12} = \boldsymbol{B}_n \boldsymbol{b}'_{n+1} + \boldsymbol{\Psi}_{12}$ with $\boldsymbol{\Psi}_{12} = \mathrm{E}(\boldsymbol{\varepsilon}_t \varepsilon_{n+1,t})$. Because of the normalisation we have $\boldsymbol{b}_{n+1}\boldsymbol{b}'_{n+1} + \sigma^2_{n+1} = 1$, where $\sigma^2_{n+1} = \mathrm{E}(\varepsilon^2_{n+1,t})$. Variable $x_{n+1,t}$ adds information if $\mathrm{E}(x_{n+1,t}\boldsymbol{x}'_t) = (\boldsymbol{b}_{n+1}\boldsymbol{B}'_n + \boldsymbol{\Psi}'_{12}) = \boldsymbol{\Gamma}'_{12} \neq 0$. This condition can be tested by means of the procedure described in Section 2.3 below.

Using the rule of determinants for partitioned matrices we get

$$\det(\boldsymbol{\Gamma}(n+1)) = \det(\boldsymbol{\Gamma}(n))(1 - a_{n+1}), \tag{7}$$

with $a_{n+1} \equiv (\boldsymbol{b}_{n+1}\boldsymbol{B}'_n + \boldsymbol{\Psi}'_{12})\boldsymbol{\Gamma}^{-1}(n)(\boldsymbol{B}_n\boldsymbol{b}'_{n+1} + \boldsymbol{\Psi}_{12})$ and $0 \leq (1 - a_{n+1}) \leq 1$. After some calculations the following relation between the relative information measures Inf^R_{n+1} and Inf^R_n can be established:

$$\mathrm{Inf}^R_{n+1} = \mathrm{Inf}^R_n - \frac{1}{n+1}\left(\frac{\log(1 - a_{n+1})}{c} + \mathrm{Inf}^R_n\right). \tag{8}$$

Therefore a variable $x_{n+1,t}$ adds relative information, i.e. $\mathrm{Inf}^R_{n+1} > \mathrm{Inf}^R_n$, if $-\log(1 - a_{n+1}) > c\mathrm{Inf}^R_n$. The second term on the right-hand side of Equation (8) serves as a threshold.

2.3 A recursive procedure

From the foregoing we have $2\mathrm{Inf}_n = -\log\det(\boldsymbol{\Gamma}(n))$ and $\mathrm{Inf}^R_n = 2\mathrm{Inf}_n/cn$.

(i) Let the first variable, i.e. the *target* variable, be $x_{1,t}$ and a collection of variables $\{x_{i,t},\ i = 2, \ldots, N\}$ with $\boldsymbol{\Gamma}(2) = \mathrm{E}\left\{\begin{pmatrix} x_{1,t} \\ x_{i,t} \end{pmatrix}\begin{pmatrix} x_{1,t} & x_{i,t} \end{pmatrix}\right\} =$

$\begin{pmatrix} 1 & r_{1,i} \\ r_{1,i} & 1 \end{pmatrix}$, where $r_{1,i}$ is the correlation between $x_{1,t}$ and $x_{i,t}$. Choose $\{x_{i,t},\ i = 2, \ldots, N\}$ such that $2\mathrm{Inf}_2 = -\log\det(\boldsymbol{\Gamma}(2)) = -\log(1 - r_{1,i}^2)$ is maximum.

(ii) From Equation (7) we have for $n = 2, 3, \ldots$

$$2\mathrm{Inf}_{n+1} = 2\mathrm{Inf}_n - \log(1 - a_{n+1}).$$

Choose the variable $\{x_{j,t},\ j = n{+}1, \ldots, N\}$ such that a_{n+1} is maximum. Then we have from Equation (8)

$$\mathrm{Inf}_{n+1}^{R,\max} = \mathrm{Inf}_n^{R,\max} - \frac{1}{n+1}\left(\frac{\log(1 - a_{n+1}^{\max})}{c} + \mathrm{Inf}_n^{R,\max}\right),$$

with increasing relative information if $a_{n+1}^{\max} > 1 - \exp(-c\mathrm{Inf}_n^{R,\max})$.

(iii) The procedure is related to Canonical Correlation (CC) and can be simplified as follows. Let $\boldsymbol{\Gamma}(n+1) = \mathrm{E}\left\{\begin{pmatrix} \boldsymbol{x}_t \\ x_{n+1,t} \end{pmatrix}\begin{pmatrix} \boldsymbol{x}_t & x_{n+1,t} \end{pmatrix}\right\} = \begin{pmatrix} \boldsymbol{\Gamma}(n) & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & 1 \end{pmatrix}$. Consider the linear transformation

$$\begin{pmatrix} \tilde{\boldsymbol{x}}_t \\ \tilde{x}_{n+1,t} \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}_1 & \mathbf{0} \\ \mathbf{0} & v^{-1} \end{pmatrix}\begin{pmatrix} \boldsymbol{x}_t \\ x_{n+1,t} \end{pmatrix},$$

with $\boldsymbol{\Gamma}(n) = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}'$ regular and $\boldsymbol{L}_1 = \boldsymbol{U}'\boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}'$ with $\boldsymbol{U}$ orthogonal, i.e. $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}' = \boldsymbol{I}_n$ and $v^2 = 1$ obtained by the SVD: $\boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}'\boldsymbol{\Gamma}_{12} =$

$\boldsymbol{U\Sigma}v$ with $\boldsymbol{\Sigma} = (\phi_{1,n+1}\ 0 \ldots 0)'$, where $\phi_{1,n+1}$ is the CC-coefficient with $0 \le \phi_{1,n+1} < 1$. The covariance of $\begin{pmatrix} \tilde{\boldsymbol{x}}_t \\ \tilde{x}_{n+1,t} \end{pmatrix}$ is $\tilde{\boldsymbol{\Gamma}}(n+1) = \begin{pmatrix} \boldsymbol{I}_n & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma}' & 1 \end{pmatrix}$. Then $2\tilde{\mathrm{Inf}}_{n+1} = -\log(1-\phi^2_{1,n+1})$ which is maximized by choosing $(x_{j,t},\ j = n+1, \ldots, N)$ such that $\phi_{1,n+1}$ is maximum, assumed to be less than one. The eigenvalues of $\tilde{\boldsymbol{\Gamma}}(n+1)$ are $\tilde{\lambda}_1 = 1 + \phi_{1,n+1}$, $\tilde{\lambda}_j = 1$ for $j = 2, \ldots, n$ and $\tilde{\lambda}_{n+1} = 1 - \phi_{1,n+1}$ and $2\tilde{\mathrm{Inf}}_{n+1}$ is maximized by minimizing the smallest eigenvalue of $\tilde{\boldsymbol{\Gamma}}(n+1)$ for $n = 3, 4, \ldots$.

The eigenvalues can be related to the Kullback-Leibler (KL) measure $2G$, see Equation (2). For $\boldsymbol{\Gamma}(n+1)$ with eigenvalues $\{\lambda_j,\ j = 1, \ldots, n+1\}$ we have $\mathrm{Inf}_{n+1} = -\sum_{j=1}^{n+1} \log \lambda_j \le \sum_{j=1}^{n+1} \frac{1}{\lambda_j} - (n+1) = G$, because $\log(x) \le x - 1$ for all positive x, and for $\tilde{\boldsymbol{\Gamma}}(n+1)$ we have $\tilde{\mathrm{Inf}}_{n+1} \le \frac{\phi^2_{1,n+1}}{1-\phi^2_{1,n+1}}$, from which it can be seen that the upper bound is maximized by choosing $\phi^2_{1,n+1}$ maximum.

Inf_{n+1} and $\tilde{\mathrm{Inf}}_{n+1}$ are related as follows. Taking determinants from

$$\tilde{\boldsymbol{\Gamma}}(n+1) = \begin{pmatrix} \boldsymbol{L}_1 & \boldsymbol{0} \\ \boldsymbol{0} & v^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Gamma}(n) & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{L}_1' & \boldsymbol{0} \\ \boldsymbol{0} & v^{-1} \end{pmatrix},$$

we have after some calculations

$$
\begin{aligned}
\det(\tilde{\boldsymbol{\Gamma}}(n+1)) =& \det(\boldsymbol{\Lambda}^{-1})\det(\boldsymbol{\Gamma}(n+1)), \quad \text{so} \\
2\mathrm{Inf}_{n+1} =& 2\mathrm{Inf}_n - \log(1-\phi^2_{1,n+1}) \qquad \text{and} \\
2\mathrm{Inf}_{n+1} =& 2\mathrm{Inf}_2 - \sum_{j=3}^{n+1} \log(1-\phi^2_{1,j}),
\end{aligned}
$$

with starting value $2\mathrm{Inf}_2 = -\log(1-r^2_{1,i})$ introduced above. Define $\delta \equiv (1-\phi^2_{1,n+1})\exp(c\mathrm{Inf}^R_n)$ we have from Equation (8) with $a_{n+1} = \phi^2_{1,n+1}$

$$
\mathrm{Inf}^R_{n+1} - \mathrm{Inf}^R_n = -\frac{1}{c(n+1)}\log\delta
$$

which is positive if $\delta < 1$, negative if $\delta > 1$, and zero if $\delta = 1$.

(iv) Replacing $\tilde{\boldsymbol{\Gamma}}(n+1)$ by a consistent estimate $\hat{\tilde{\boldsymbol{\Gamma}}}(n+1)$ and applying the same SVD procedure yields $\hat{\tilde{\mathrm{Inf}}}_{n+1} = -\log(1-\hat{\tilde{\phi}}^2_1)/2$. Under $H_0:\ \phi_1 = 0$, the Bartlett test statistic

$$
-[T-1/2(n+2)]\log(1-\hat{\tilde{\phi}}^2_1) = [T-1/2(n+2)]2\hat{\tilde{\mathrm{Inf}}}_{n+1}
$$

follows asymptotically a χ^2-distribution with n degrees of freedom, see e.g. Muirhead [16]. Testing the hypothesis $\phi_1 = 0$ is basically testing whether the transformed vector $(\tilde{\boldsymbol{x}}'_t\ \tilde{x}_{n+1,t})'$ has maximum entropy, i.e. no correlation at all.

2.4 MSE-prediction

From the foregoing we have $\tilde{\boldsymbol{x}}_t = \boldsymbol{L}_1 \boldsymbol{x}_t$ with $\boldsymbol{L}_1 = \boldsymbol{U}'\boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}'$. Given a realization $\tilde{x}_{n+1,t} = v^{-1}x_{n+1,t}$ the conditional mean (predictor) of $\tilde{\boldsymbol{x}}_t$ is $\tilde{\boldsymbol{x}}_t^P = \boldsymbol{\Sigma}\tilde{x}_{n+1,t}$ with conditional variance $\text{var}\{\tilde{\boldsymbol{x}}_t^P\} = \boldsymbol{I} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}' = \text{diag}((1 - \phi_1^2),\ 1, \ldots, 1)$ and information $-\log(1-\phi_1^2)/2$. Hence if $\phi_1 = 0$ implying $\boldsymbol{\Sigma} = 0$ the vector $\tilde{\boldsymbol{x}}_t^P$ has maximum entropy and no information.

The conditional MSE-predictor of $\boldsymbol{x}_t$ itself is

$$\boldsymbol{x}_t^P = \boldsymbol{L}_1^{-1}\tilde{\boldsymbol{x}}_t^P = \phi_1 \boldsymbol{C}\boldsymbol{\Lambda}^{1/2}\boldsymbol{u}_1 \tilde{x}_{n+1,t},$$

where $\boldsymbol{u}_1$ is the first column of the orthonormal matrix $\boldsymbol{U}$. The conditional variance of $\boldsymbol{x}_t^P$ is

$$\text{var}\{\boldsymbol{x}_t^P\} = \boldsymbol{L}_1^{-1}\left(\boldsymbol{L}_1^{-1}\right)' - \boldsymbol{L}_1^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}'\left(\boldsymbol{L}_1^{-1}\right)' = \boldsymbol{\Gamma}(n) - \boldsymbol{L}_1^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}'\left(\boldsymbol{L}_1^{-1}\right)',$$

from which it can be seen that $\boldsymbol{\Gamma}(n)$ exceeds $\text{var}\{\boldsymbol{x}_t^P\}$ by a positive definite matrix if $\phi_1 > 0$. Therefore adding a variable in case $\phi_1 > 0$ increases the MSE prediction quality measured as a decrease in the variance of $\boldsymbol{x}_t^P$.

2.5 Comparison to standard information criterion-based measures

Let $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{X}^* \end{bmatrix} \in \mathbb{R}^{T\times N}$ with $\boldsymbol{x}_1 \in \mathbb{R}^T$ the time series of the target variable $x_{1,t}$ and $\boldsymbol{X}^*$ the ordered data set according to the procedure described above. Apply a Singular Value Decomposition (SVD)

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{C}' = \boldsymbol{U}_1\boldsymbol{S}_1\boldsymbol{C}_1' + \boldsymbol{U}_2\boldsymbol{S}_2\boldsymbol{C}_2' = \hat{\boldsymbol{X}} + \boldsymbol{E}, \tag{9}$$

where $\boldsymbol{U}_1 \in \mathbb{R}^{T\times k}$ consists of the first k principal components (PC) of $\boldsymbol{X}$. This procedure is identical to Stock and Watson [20], who propose principal components as an estimator for unobserved factors $\boldsymbol{F}^{SW}$ and, subsequently, employ a linear projection of the data on the factors to estimate the factor loadings. The largest k eigenvectors of the sample covariance matrix $\frac{1}{T}\boldsymbol{X'X}$ can be obtained as $\frac{1}{\sqrt{T}}\boldsymbol{X'}\frac{1}{\sqrt{T}}\boldsymbol{X'} = \boldsymbol{C}_1\bar{\boldsymbol{S}}_1^2\boldsymbol{C}_1'$ and so, in matrix notation, $\widehat{\boldsymbol{F}}^{SW} = \boldsymbol{X}\boldsymbol{C}_1$. Let the factor loadings matrix $\boldsymbol{B}^{SW}$ be obtained by the linear projection of $\boldsymbol{X}$ on $\widehat{\boldsymbol{F}}^{SW}$. So, $\boldsymbol{X} = \widehat{\boldsymbol{F}}^{SW}\boldsymbol{B}^{SW}$ and $\widehat{\boldsymbol{B}}^{SW} = \left(\widehat{\boldsymbol{F}}^{SW'}\widehat{\boldsymbol{F}}^{SW}\right)^{-1}\widehat{\boldsymbol{F}}^{SW'}\boldsymbol{X}$ that leads to $\widehat{\boldsymbol{B}}^{SW} = \boldsymbol{C}_1'$ and $\widehat{\boldsymbol{X}}^{SW} = \boldsymbol{X}\boldsymbol{C}_1\boldsymbol{C}_1'$. To see the equivalence, employ Equation (9) for $\boldsymbol{X}$, which leads to $\widehat{\boldsymbol{X}}^{SW} = \left[\boldsymbol{U}_1\boldsymbol{S}_1\boldsymbol{C}_1' + \boldsymbol{U}_2\boldsymbol{S}_2\boldsymbol{C}_2'\right]\boldsymbol{C}_1\boldsymbol{C}_1' = \boldsymbol{U}_1\boldsymbol{S}_1\boldsymbol{C}_1'$, or equivalently, $\widehat{\boldsymbol{F}}^{SW} = \boldsymbol{X}\boldsymbol{C}_1 = \boldsymbol{U}_1\boldsymbol{S}_1$.

We partition

$$\begin{aligned}
\hat{\boldsymbol{X}} &= \begin{bmatrix} \hat{\boldsymbol{x}}_1 & \hat{\boldsymbol{X}}^* \end{bmatrix} = \boldsymbol{U}_1\boldsymbol{S}_1 \begin{bmatrix} \bar{c}_{11} & \bar{\boldsymbol{C}}_{12} \end{bmatrix} \quad \text{and} \\
\boldsymbol{E} &= \begin{bmatrix} \boldsymbol{e}_{\hat{x}_1} & \boldsymbol{E}_{\hat{\boldsymbol{X}}^*} \end{bmatrix} = \boldsymbol{U}_2\boldsymbol{S}_2 \begin{bmatrix} \bar{c}_{21} & \bar{\boldsymbol{C}}_{22} \end{bmatrix}.
\end{aligned}$$

The PC estimate of $\boldsymbol{x}_1$ is $\hat{\boldsymbol{x}}_1 = \boldsymbol{U}_1\boldsymbol{S}_1\bar{c}_{11}$ with error $\boldsymbol{e}_{\hat{x}_1} = \boldsymbol{U}_2\boldsymbol{S}_2\bar{c}_{21}$ and $\hat{\boldsymbol{x}}_1'\hat{\boldsymbol{x}}_1 = \sum_{j=1}^{k} s_j^2\bar{c}_{11,j}^2$ and $\boldsymbol{e}_{\hat{x}_1}'\boldsymbol{e}_{\hat{x}_1} = \sum_{j=k+1}^{N} s_j^2\bar{c}_{21,j-k}^2$. Since $\boldsymbol{x}_1$ is standardized, it holds that $(\hat{\boldsymbol{x}}_1'\hat{\boldsymbol{x}}_1 + \boldsymbol{e}_{\hat{x}_1}'\boldsymbol{e}_{\hat{x}_1})/T = 1$ and so, we can interpret the commonality ratio $\hat{\boldsymbol{x}}_1'\hat{\boldsymbol{x}}_1/T$ as the part of the variance that can be approximated by using the factor basis $\boldsymbol{U}_1\boldsymbol{S}_1$.

The Akaike information criterion (AIC) for this model (see e.g. Greene [11] Section 7.4) becomes

$$\mathrm{AIC}(k) = \log\left(\sum_{j=k+1}^{N} \hat{\lambda}_j \bar{c}^2_{21,j-k}\right) + \frac{2k}{T}, \tag{10}$$

where $\boldsymbol{X}'\boldsymbol{X}/T = \boldsymbol{C}\hat{\boldsymbol{\Lambda}}\boldsymbol{C}'$ with $s_j^2/T = \hat{\lambda}_j$ and $T > N$. The quality of the selection procedure can be judged with the AIC of Equation (10) for increasing number of variables n.

3 Monte Carlo experiment

We generate data from the generalized dynamic factor structure

$$x_{it} = B_{i1}(L)\,F_{1t} + ... + B_{k1}(L)\,F_{kt} + e_{it}, \tag{11}$$

where $B_{i1}(L) = \sum_{i=0}^{\infty} B_{ij}^{(u)} L^u$ with lag operator L, factor loadings $B_{ij}^{(u)}$, factors F_{jt} and idiosyncratic term e_{it}. We replicate Onatski's [17] modification of Hallin and Liška's [12] Monte Carlo experiment and generate data from model (11) as follows:

1. the k-dimensional factor vectors F_{jt} are i.i.d. $N(0, I_k)$.

2. the filters $B_{ik}(L)$, $(i = 1, ..., n;\ k = 1, ..., q)$ are randomly generated independently from the F_{jt}'s by the AR loadings: $B_{ik}(L) =$

$b_{ij}^{(0)}\left(1-b_{ij}^{(1)}L\right)^{-1}\left(1-b_{ij}^{(2)}L\right)^{-1}$ with i.i.d. and mutually independent coefficients $b_{ij}^{(0)}\sim\mathcal{N}(0,1)$, $b_{ij}^{(1)}\sim\mathcal{U}[.8,.9]$ and $b_{ij}^{(2)}\sim\mathcal{U}[.5,.6]$

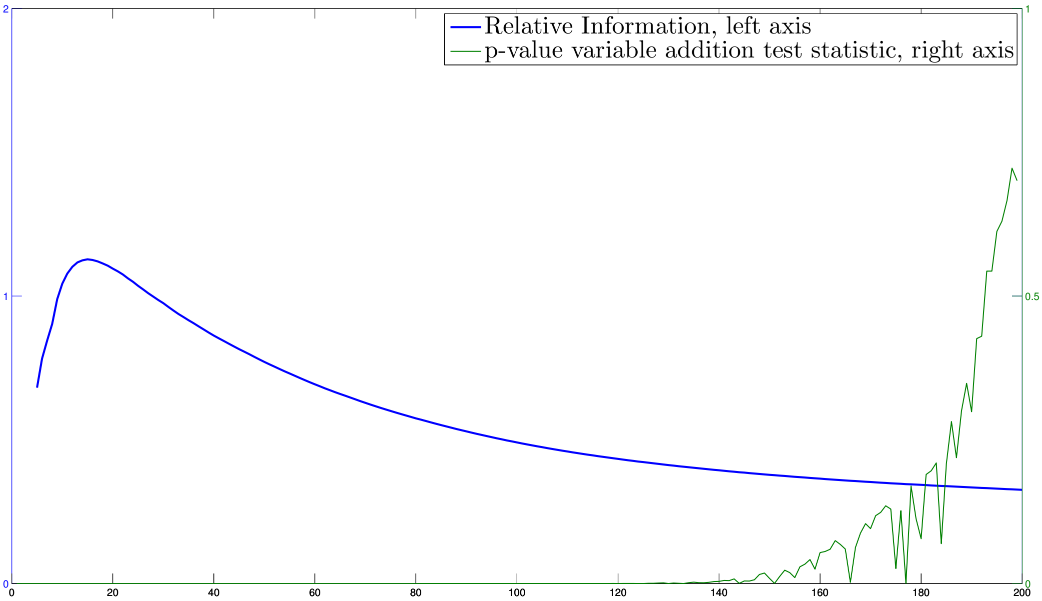
3. the idiosyncratic components e_{it} follow $AR(1)$ processes both cross-sectionally and over time: $e_{it}=\rho_i e_{it-1}+v_{it}$ and $v_{it}=\rho v_{i-1t}+u_{it}$, with i.i.d coefficients $\rho_i\sim\mathcal{U}[-.8,.8]$, $\rho=0.2$ and $u_{it}\sim\mathcal{N}(0,1)$ i.i.d. and independently generated from $B_{ik}(L)$ and F_{jt}, cf. Onatski [17]. The support $[-.8,.8]$ of the uniform distribution has been chosen to match the range of the first-order autocorrelations of the estimated idiosyncratic components of the Stock and Watson [21] dataset.

4. For each i, the variance of e_{it} and that of the common components $\sum_{j=1}^{k}B_{ij}(L)F_{jt}$ are normalized such that their variances equal $0.4+0.05k$ and $1-(0.4+0.05k)$, respectively. Hence, a $2-$factor model explains 50% of the data variation and a $7-$factor model 75% for $\sigma=1$. As a final step, the idiosyncratic part is magnified by $\sigma\geq 1$.

We calibrate the Monte Carlo simulation with $T=500$, $N=200$, $k=3$, $\sigma=3$, $\rho=0.2$ and finally, we magnify the idiosyncratic part by i/N and the common part by $(N-i)/N$ for $i=1,...,N$. Then, we implement the recursive procedure of Section 2.3 using the first generated variable of the simulation as the target variable. Figure 1 shows the relative information criterion and the p-values of the variable addition test statistic. Figure 2 shows the corresponding commonality ratio $\hat{\boldsymbol{x}}_1'\hat{\boldsymbol{x}}_1/T$, and the AIC criterion of Section 2.5. For both figures, the ordered data set runs from $n=4,...,N$ to ensure that the number of variables is larger than the number of factors

$k = 3$. Note that the second term in Equation (10) is a constant, when the number of factors k is fixed.

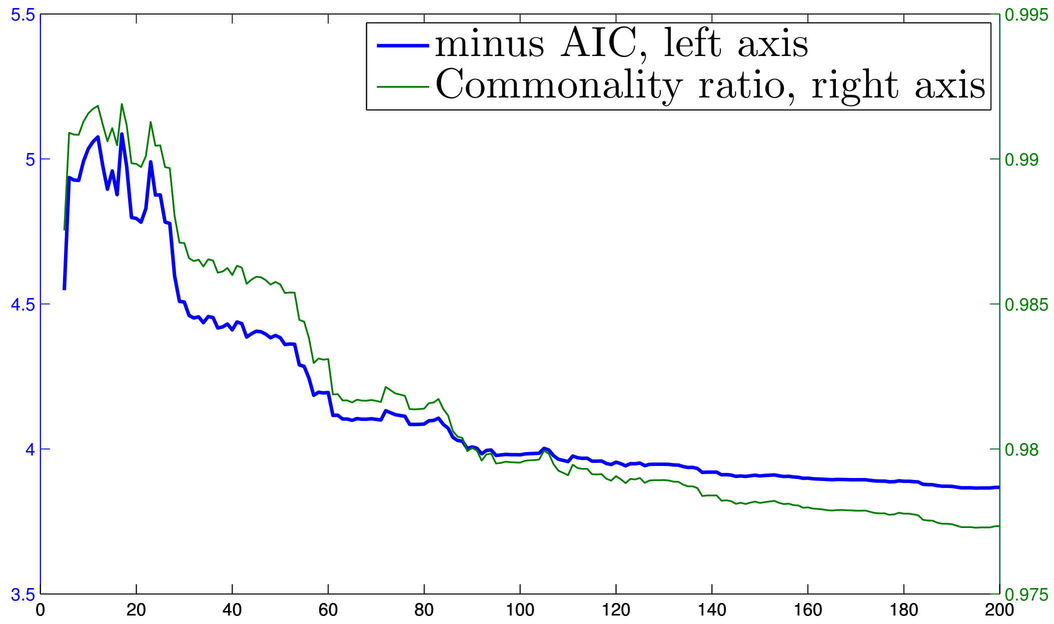
Figure 1: Relative information



The p-value of the variable addition test in Figure 1 indicates that a lot of series are informative, whereas the relative information—measured by the ratio of information, Inf_N, and maximum entropy cN —is maximized for around 20 series. The latter observation also holds for the commonality ratio and the AIC of Figure 2, computed for $k = 3$. More than this number of series add information to the ordered data set, i.e. $\text{Inf}_{N+1} > \text{Inf}_N$, but apparently the additional information does not exceed the increase in entropy in these series, $\text{Inf}_{N+1} - \text{Inf}_N < c(N + 1) - cN = c$, and therefore $\text{Inf}^R_{n+1} < \text{Inf}^R_n$.

We conclude that our relative information measure, the commonality ratio and the AIC lead to a selection of around 20 series. However, unlike the other two statistics, our measure does not rely on a priori knowing the number of factors.

Figure 2: AIC and commonality ratio



4 Application

In the application below, we use the relative information measure introduced above to order a macroeconomic data set. Plots of the relative information measures against the number of variables indicate which subset is most informative for factor modelling.

4.1 The Stock and Watson data set

In this section we evaluate the performance of the suggested approach on the U.S. macroeconomic data set of [21], which consists of monthly observations on $N = 132$ macroeconomic time series from 1959M1 up to and including 2003M12 ($T = 540$). The series cover 14 categories: real output and income; employment and hours; real retail, manufacturing and trade sales; consumption; housing starts and sales; real inventories; orders; stock prices; exchange rates; interest rates and spreads; money and credit quantity aggregates; price indexes; average hourly earnings; and miscellaneous. The series are transformed by taking logarithms and/or differencing when necessary to assure approximate stationarity. In general, first differences of logarithms (growth rates) are used for real quantity variables, first differences are used for nominal interest rates, and second differences of logarithms for price series (changes in inflation). Moreover, the series are adjusted for outliers by replacing the observations of the transformed variables with absolute median deviations larger than 6 times the interquartile range with the median value of the preceding 5 observations. The specific transformations and the list of series are given in Appendix A of Stock and Watson [21].

Concerning the number of factors to represent the data set, different test procedures are proposed and employed. For instance, Hallin and Liška [12] find $\widehat{k} = 1$ factor for the whole sample, but $\widehat{k} = 3$ factors for the period 1960–1982. Onatski [17] restricts the analysis to business cycle frequencies and explicitly excludes cycles longer than 10 years. Employing his test procedure as an algorithm procedure results in $\widehat{k} = 1$ factors. Bai and Ng [3] estimate

$\widehat{k} = 4$ factors, but point out that there is substantial variation over the sample. Finally, Otter, Jacobs and den Reijer [19] also find $\widehat{k} = 1$ for the whole sample and substantial variation for the first part. In the computation of the AIC and the commonality ratio below, the number of factors is set to $\widehat{k} = 3$. This choice does not affect the relative information outcomes which are based on the recursive procedure of Section 2.3.

4.2 Information in the data set

Using the recursive procedure described in Section 2.3, we order the data set according to the relative information measure with respect to two target variables: the first difference of the log of total industrial production (IP hereafter) and the second difference of the log of the consumer price index (CPI hereafter)The full data set consists of $N = 132$ time series variables, with $T = 540$ observations covering the sample 1959M1–2003M12. Since the number of observations T is much larger than the number of series N, all eigenvalues of the covariance matrix of $\boldsymbol{x}_t$ differ from zero and our relative information measure is computationally stable.

Table 1 presents the orders of the first 50 variables according to the two relative information criteria for both target variables. The table allows the following observations. The first ten series that are included in the subset for IP belong to the group of Industrial Production; the first ten series for CPI are price indices. Second, price indices are generally speaking not informative for IP (the exception is series # 114: NAPM commodity price index), while production series do not appear in the first fifty variables of the ordered data subset for CPI (with one exception series # 19: NAPM production). Finally, variables enter the ordered data sets in clusters. For IP, the relative information measure first selects a group of industrial production variables, followed by employment series, interest rates and spreads, and housing starts and sales. With CPI as target variable, the relative information measure starts with picking price indices, followed by employment, orders, interest rates and spreads, housing starts and sales, and employment.

Table 1: Ranking of series according to relative information

order	IP series #	CPI series #
1	6	115
2	16	124
3	20	123
4	7	119
5	8	125
6	13	127
7	14	122
8	9	117
9	12	128
10	11	121
11	19	39
12	62	37
13	61	38
14	50	34
15	64	33
16	37	40
17	38	41
18	34	43
19	33	50
20	40	61
21	41	19
22	43	62
23	42	64
24	63	42
25	114	63
26	39	114
27	102	102
28	101	101
29	100	100
30	99	99
31	97	98
32	96	97
33	98	96
34	95	95
35	59	59
36	54	54
37	56	56
38	51	51
39	60	60
40	55	55
41	58	58
42	53	53
43	57	57
44	52	52
45	49	49
46	47	47
47	44	44
48	36	36
49	74	74
50	68	68

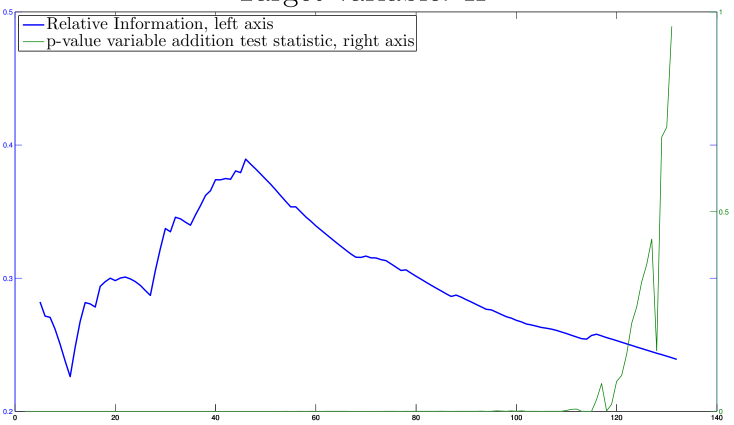
Notes. See the table in the appendix for the description of the variables.

Figure 3 shows the evolution in relative information if we order the data set according to the target variables IP (top panel) and CPI (bottom panel). The figure reveals that sometimes relative information, or weighted mean information per variable, decreases with the addition of a single series, but increases if a batch of variables is added. For both target variables relative information attains a global maximum if we take between 40 and 50 series in line with the findings of Boivin and Ng [6] and Inklaar et al. [13]. This conclusion is supported by the AIC and commonality ratio in Figure 4.

Figure 3 also shows p-values of the test described in Section 2.3 whether an additional variable adds information. The null hypothesis is that an additional variable is not correlated with the variables already included in the set. Hence, low p-values indicate that an additional variable adds information. We note that the outcomes of the test are not sensitive to the initial condition, i.e. the choice of the target variable. The figure suggest that some 120 series are informative. This finding does not contradict our conclusion that *relative* information, or weighted mean information per variable, measured by the ratio of information, Inf_N, and maximum entropy cN, is maximized for 40–50 series. More than this number of series add *absolute* information to the ordered data set, i.e. $\mathrm{Inf}_{N+1} > \mathrm{Inf}_N$ for $40 < N < 120$, because $\log(1 - a_{N+1}) < 0$ for $0 < a_{N+1} < 1$, see Equation (7). However, for $40 < N < 120$ we have for the relative information measure $c\mathrm{Inf}_N^R > |\log(1 - a_{N+1})|$, see Equation (8), cf the discussion at the end of Section 3.

Figure 3: Relative information of ordered data set

Target variable: IP



Target variable: CPI

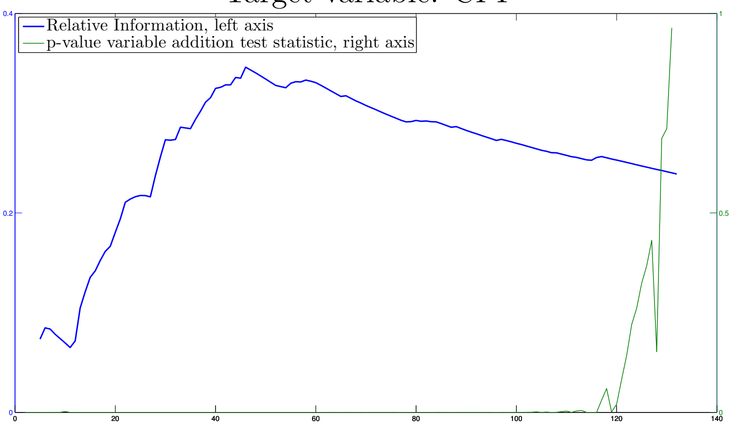
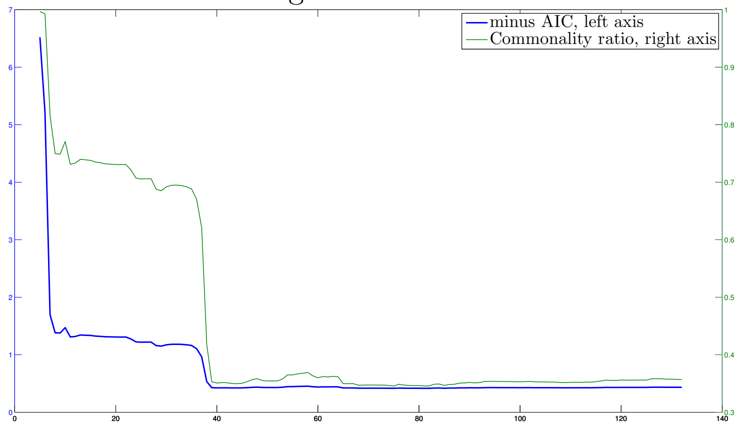


Figure 4: AIC and commonality ratio

Target variable: IP



Target variable: CPI

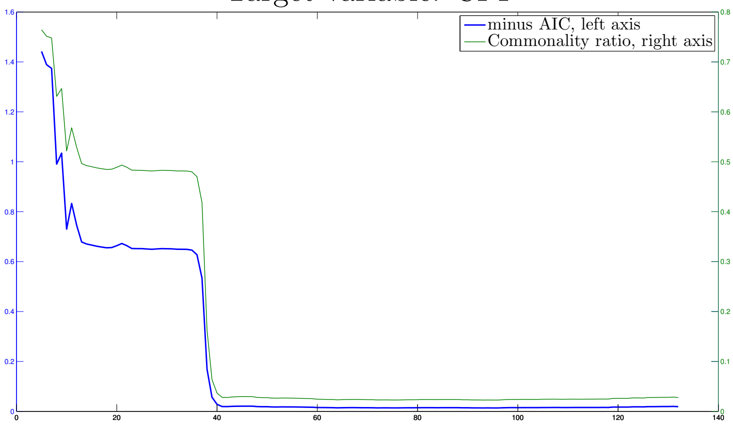


Figure 4 shows the corresponding AIC and commonality ratio for fixed $k = 3$, and the increasing number of series N. Our relative information measure has an optimum at around 40 chosen variables, which coincides with nearly constant values of the AIC and the commonality ratio. This suggests homogeneity of the first 40 chosen variables according to our method.

5 Conclusion

This paper fruitfully applied concepts from information theory in the analysis of large data sets. We defined a relative information measure linked to Kullback-Leibler numbers. The application of the measures enabled us to order a data set and to identify a subset of the data that is most informative.

We illustrated our methods with a Monte Carlo study and the Stock and Watson U.S. macroeconomic data set consisting of 132 times series variables with 540 observations. Both analyses show that relative information is maximized for a limited number of series. In the Stock and Watson data set relative information is maximized for around 40–50 series if we are interested in modelling industrial production and CPI inflation. We conclude that our method can indeed produce a considerable reduction in the dimension of a data set, which implies less series that have to be monitored.

Our relative information measure is based on the eigenvalues of the covariance matrix of the data, which is only defined if the number of observations T exceeds the number of series N. Future research will deal with the mirror situation of $N > T$.

Appendix A: The Stock and Watson U.S. macroeconomic data set

Table A.1 lists the 132 series of the Stock and Watson [21] U.S. data set, with number, mnemonic, and description of the variable. For details like the transformation applied to the series and sources see Stock and Watson [21] Appendix A. As is required for factor estimation, the variables are standardized by subtracting their mean and then dividing by their standard deviation. This standardization is necessary to avoid overweighting of large variance series in the factor estimation.

Table A.1: Description of the Stock and Watson data set

#	Short name	Mnemonic	Description
1	PI	A0M052	Personal income (AR, bil. chain 2000 $)
2	PI less transfers	A0M051	Personal income less transfer payments (AR, bil. chain 2000 $)
3	Consumption	A0M224_R	Real Consumption (AC) A0m224/gmdc
4	M&T sales	A0M057	Manufacturing and trade sales (mil. Chain 1996 $)
5	Retail sales	A0M059	Sales of retail stores (mil. Chain 2000 $)
6	IP: total	IPS10	INDUSTRIAL PRODUCTION INDEX - TOTAL INDEX
7	IP: products	IPS11	INDUSTRIAL PRODUCTION INDEX - PRODUCTS, TOTAL
8	IP: final prod	IPS299	INDUSTRIAL PRODUCTION INDEX - FINAL PRODUCTS
9	IP: cons gds	IPS12	INDUSTRIAL PRODUCTION INDEX - CONSUMER GOODS
10	IP: cons dble	IPS13	INDUSTRIAL PRODUCTION INDEX - DURABLE CONSUMER GOODS
11	iIP:cons nondble	IPS18	INDUSTRIAL PRODUCTION INDEX - NONDURABLE CONSUMER GOODS
12	IP:bus eqpt	IPS25	INDUSTRIAL PRODUCTION INDEX - BUSINESS EQUIPMENT
13	IP: matls	IPS32	INDUSTRIAL PRODUCTION INDEX - MATERIALS
14	IP: dble mats	IPS34	INDUSTRIAL PRODUCTION INDEX - DURABLE GOODS MATERIALS
15	IP:nondble mats	IPS38	INDUSTRIAL PRODUCTION INDEX - NONDURABLE GOODS MATERIALS
16	IP: mfg	IPS43	INDUSTRIAL PRODUCTION INDEX - MANUFACTURING (SIC)
17	IP: res util	IPS307	INDUSTRIAL PRODUCTION INDEX - RESIDENTIAL UTILITIES
18	IP: fuels	IPS306	INDUSTRIAL PRODUCTION INDEX - FUELS
19	NAPM prodn	PMP	NAPM PRODUCTION INDEX (PERCENT)
20	Cap util	A0M082	Capacity Utilization (Mfg)
21	Help wanted indx	LHEL	INDEX OF HELP-WANTED ADVERTISING IN NEWSPAPERS (1967=100;SA)
22	Help wanted/emp	LHELX	EMPLOYMENT: RATIO; HELP-WANTED ADS:NO. UNEMPLOYED CLF
23	Emp CPS total	LHEM	CIVILIAN LABOR FORCE: EMPLOYED, TOTAL (THOUS.,SA)
24	Emp CPS nonag	LHNAG	CIVILIAN LABOR FORCE: EMPLOYED, NONAGRIC.INDUSTRIES (THOUS.,SA)
25	U: all	LHUR	UNEMPLOYMENT RATE: ALL WORKERS, 16 YEARS & OVER (%,SA)
26	U: mean duration	LHU680	UNEMPLOY.BY DURATION: AVERAGE(MEAN)DURATION IN WEEKS (SA)
27	U ¡ 5 wks	LHU5	UNEMPLOY.BY DURATION: PERSONS UNEMPL.LESS THAN 5 WKS (THOUS.,SA)
28	U 5-14 wks	LHU14	UNEMPLOY.BY DURATION: PERSONS UNEMPL.5 TO 14 WKS (THOUS.,SA)
29	U 15+ wks	LHU15	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 WKS + (THOUS.,SA)
30	U 15-26 wks	LHU26	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 TO 26 WKS (THOUS.,SA)
31	U 27+ wks	LHU27	UNEMPLOY.BY DURATION: PERSONS UNEMPL.27 WKS + (THOUS,SA)
32	UI claims	A0M005	Average weekly initial claims, unemploy. insurance (thous.)
33	Emp: total	CES002	EMPLOYEES ON NONFARM PAYROLLS - TOTAL PRIVATE
34	Emp: gds prod	CES003	EMPLOYEES ON NONFARM PAYROLLS - GOODS-PRODUCING
35	Emp: mining	CES006	EMPLOYEES ON NONFARM PAYROLLS - MINING
36	Emp: const	CES011	EMPLOYEES ON NONFARM PAYROLLS - CONSTRUCTION
37	Emp: mfg	CES015	EMPLOYEES ON NONFARM PAYROLLS - MANUFACTURING
38	Emp: dble gds	CES017	EMPLOYEES ON NONFARM PAYROLLS - DURABLE GOODS
39	Emp: nondbles	CES033	EMPLOYEES ON NONFARM PAYROLLS - NONDURABLE GOODS
40	Emp: services	CES046	EMPLOYEES ON NONFARM PAYROLLS - SERVICE-PROVIDING
41	Emp: TTU	CES048	EMPLOYEES ON NONFARM PAYROLLS - TRADE, TRANSPORTATION, AND UTILITIES
42	Emp: wholesale	CES049	EMPLOYEES ON NONFARM PAYROLLS - WHOLESALE TRADE
43	Emp: retail	CES053	EMPLOYEES ON NONFARM PAYROLLS - RETAIL TRADE
44	Emp: FIRE	CES088	EMPLOYEES ON NONFARM PAYROLLS - FINANCIAL ACTIVITIES
45	Emp: Govt	CES140	EMPLOYEES ON NONFARM PAYROLLS - GOVERNMENT
46	Emp-hrs nonag	A0M048	Employee hours in nonag. establishments (AR, bil. hours)
47	Avg hrs	CES151	AVERAGE WEEKLY HOURS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFARM PAYROLLS - GOODS-PRODUCING
48	Overtime: mfg	CES155	AVERAGE WEEKLY HOURS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFARM PAYROLLS - MFG OVERTIME HOURS
49	Avg hrs: mfg	AOM001	Average weekly hours, mfg. (hours)
50	NAPM empl	PMEMP	NAPM EMPLOYMENT INDEX (PERCENT)
51	HStarts: Total	HSFR	HOUSING STARTS:NONFARM(1947-58);TOTAL FARM&NONFARM(1959-)(THOUS.,SAAR)
52	HStarts: NE	HSNE	HOUSING STARTS:NORTHEAST (THOUS.U.)S.A.
53	HStarts: MW	HSMW	HOUSING STARTS:MIDWEST(THOUS.U.)S.A.
54	HStarts: South	HSSOU	HOUSING STARTS:SOUTH (THOUS.U.)S.A.
55	HStarts: West	HSWST	HOUSING STARTS:WEST (THOUS.U.)S.A.
56	BP: total	HSBR	HOUSING AUTHORIZED: TOTAL NEW PRIV HOUSING UNITS (THOUS.,SAAR)
57	BP: NE	HSBNE	HOUSES AUTHORIZED BY BUILD. PERMITS:NORTHEAST(THOU.U.)S.A
58	BP: MW	HSBMW	HOUSES AUTHORIZED BY BUILD. PERMITS:MIDWEST(THOU.U.)S.A.
59	BP: South	HSBSOU	HOUSES AUTHORIZED BY BUILD. PERMITS:SOUTH(THOU.U.)S.A.
60	BP: West	HSBWST	HOUSES AUTHORIZED BY BUILD. PERMITS:WEST(THOU.U.)S.A.
61	PMI	PMI	PURCHASING MANAGERS' INDEX (SA)
62	NAPM new ordrs	PMNO	NAPM NEW ORDERS INDEX (PERCENT)
63	NAPM vendor del	PMDEL	NAPM VENDOR DELIVERIES INDEX (PERCENT)
64	NAPM Invent	PMNV	NAPM INVENTORIES INDEX (PERCENT)

#	Short name	Mnemonic	Description
65	Orders: cons gds	A0M008	Mfrs' new orders, consumer goods and materials (bil. chain 1982 $)
66	Orders: dble gds	A0M007	Mfrs' new orders, durable goods industries (bil. chain 2000 $)
67	Orders: cap gds	A0M027	Mfrs' new orders, nondefense capital goods (mil. chain 1982 $)
68	Unf orders: dble	A1M092	Mfrs' unfilled orders, durable goods indus. (bil. chain 2000 $)
69	M&T invent	A0M070	Manufacturing and trade inventories (bil. chain 2000 $)
70	M&T invent/sales	A0M077	Ratio, mfg. and trade inventories to sales (based on chain 2000 $)
71	M1	FM1	MONEY STOCK: M1(CURR,TRAV.CKS,DEM DEP,OTHER CK'ABLE DEP)(BIL$,SA)
72	M2	FM2	MONEY STOCK:M2(M1+O'NITE RPS,EURO$,G/P&B/D MMMFS&SAV&SM TIME DEP(BIL$,SA)
73	M3	FM3	MONEY STOCK: M3(M2+LG TIME DEP,TERM RP'S&INST ONLY MMMFS)(BIL$,SA)
74	M2 (real)	FM2DQ	MONEY SUPPLY - M2 IN 1996 DOLLARS (BCI)
75	MB	FMFBA	MONETARY BASE, ADJ FOR RESERVE REQUIREMENT CHANGES(MIL$,SA)
76	Reserves tot	FMRRA	DEPOSITORY INST RESERVES:TOTAL,ADJ FOR RESERVE REQ CHGS(MIL$,SA)
77	Reserves nonbor	FMRNBA	DEPOSITORY INST RESERVES:NONBORROWED,ADJ RES REQ CHGS(MIL$,SA)
78	C&I loans	FCLNQ	COMMERCIAL & INDUSTRIAL LOANS OUSTANDING IN 1996 DOLLARS (BCI)
79	C&I loans	FCLBMC	WKLY RP LG COM'L BANKS:NET CHANGE COM'L & INDUS LOANS(BIL$,SAAR)
80	Cons credit	CCINRV	CONSUMER CREDIT OUTSTANDING - NONREVOLVING(G19)
81	Inst cred/PI	A0M095	Ratio, consumer installment credit to personal income (pct.)
82	S&P 500	FSPCOM	S&P'S COMMON STOCK PRICE INDEX: COMPOSITE (1941-43=10)
83	S&P: indust	FSPIN	S&P'S COMMON STOCK PRICE INDEX: INDUSTRIALS (1941-43=10)
84	S&P div yield	FSDXP	S&P'S COMPOSITE COMMON STOCK: DIVIDEND YIELD (% PER ANNUM)
85	S&P PE ratio	FSPXE	S&P'S COMPOSITE COMMON STOCK: PRICE-EARNINGS RATIO (%,NSA)
86	FedFunds	FYFF	INTEREST RATE: FEDERAL FUNDS (EFFECTIVE) (% PER ANNUM,NSA)
87	Commpaper	CP90	Commercial Paper Rate (AC)
88	3 mo T-bill	FYGM3	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,3-MO.(% PER ANN,NSA)
89	6 mo T-bill	FYGM6	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,6-MO.(% PER ANN,NSA)
90	1 yr T-bond	FYGT1	INTEREST RATE: U.S.TREASURY CONST MATURITIES,1-YR.(% PER ANN,NSA)
91	5 yr T-bond	FYGT5	INTEREST RATE: U.S.TREASURY CONST MATURITIES,5-YR.(% PER ANN,NSA)
92	10 yr T-bond	FYGT10	INTEREST RATE: U.S.TREASURY CONST MATURITIES,10-YR.(% PER ANN,NSA)
93	Aaabond	FYAAAC	BOND YIELD: MOODY'S AAA CORPORATE (% PER ANNUM)
94	Baa bond	FYBAAC	BOND YIELD: MOODY'S BAA CORPORATE (% PER ANNUM)
95	CP-FF spread	SCP90	cp90-fyff
96	3 mo-FF spread	SFYGM3	fygm3-fyff
97	6 mo-FF spread	SFYGM6	fygm6-fyff
98	1 yr-FF spread	SFYGT1	fygt1-fyff
99	5 yr-FFspread	SFYGT5	fygt5-fyff
100	10yr-FF spread	SFYGT10	fygt10-fyff
101	Aaa-FF spread	SFYAAAC	fyaaac-fyff
102	Baa-FF spread	SFYBAAC	fybaac-fyff
103	Ex rate: avg	EXRUS	UNITED STATES;EFFECTIVE EXCHANGE RATE(MERM)(INDEX NO.)
104	Ex rate: Switz	EXRSW	FOREIGN EXCHANGE RATE: SWITZERLAND (SWISS FRANC PER U.S.$)
105	Ex rate: Japan	EXRJAN	FOREIGN EXCHANGE RATE: JAPAN (YEN PER U.S.$)
106	Ex rate: UK	EXRUK	FOREIGN EXCHANGE RATE: UNITED KINGDOM (CENTS PER POUND)
107	EX rate: Canada	EXRCAN	FOREIGN EXCHANGE RATE: CANADA (CANADIAN $ PER U.S.$)
108	PPI: fin gds	PWFSA	PRODUCER PRICE INDEX: FINISHED GOODS (82=100,SA)
109	PPI: cons gds	PWFCSA	PRODUCER PRICE INDEX:FINISHED CONSUMER GOODS (82=100,SA)
110	PPI: int matls	PWIMSA	PRODUCER PRICE INDEX:INTERMED MAT.SUPPLIES & COMPONENTS(82=100,SA)
111	PPI: crude matls	PWCMSA	PRODUCER PRICE INDEX:CRUDE MATERIALS (82=100,SA)
112	Commod: spot price	PSCCOM	SPOT MARKET PRICE INDEX:BLS & CRB: ALL COMMODITIES(1967=100)
113	Sens matls price	PSM99Q	INDEX OF SENSITIVE MATERIALS PRICES (1990=100)(BCI-99A)
114	NAPM com price	PMCP	NAPM COMMODITY PRICES INDEX (PERCENT)
115	CPI-U: all	PUNEW	CPI-U: ALL ITEMS (82-84=100,SA)
116	CPI-U: apparel	PU83	CPI-U: APPAREL & UPKEEP (82-84=100,SA)
117	CPI-U: transp	PU84	CPI-U: TRANSPORTATION (82-84=100,SA)
118	CPI-U: medical	PU85	CPI-U: MEDICAL CARE (82-84=100,SA)
119	CPI-U: comm.	PUC	CPI-U: COMMODITIES (82-84=100,SA)
120	CPI-U: dbles	PUCD	CPI-U: DURABLES (82-84=100,SA)
121	CPI-U: services	PUS	CPI-U: SERVICES (82-84=100,SA)
122	CPI-U: ex food	PUXF	CPI-U: ALL ITEMS LESS FOOD (82-84=100,SA)
123	CPI-U: ex shelter	PUXHS	CPI-U: ALL ITEMS LESS SHELTER (82-84=100,SA)
124	CPI-U: ex med	PUXM	CPI-U: ALL ITEMS LESS MEDICAL CARE (82-84=100,SA)
125	PCE defl	GMDC	PCE,IMPL PR DEFL:PCE (1987=100)
126	PCE defl: dlbes	GMDCD	PCE,IMPL PR DEFL:PCE; DURABLES (1987=100)
127	PCE defl: nondble	GMDCN	PCE,IMPL PR DEFL:PCE; NONDURABLES (1996=100)
128	PCE defl: services	GMDCS	PCE,IMPL PR DEFL:PCE; SERVICES (1987=100)
129	AHE: goods	CES275	AVERAGE HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFARM PAYROLLS - GOODS PRODUCING
130	AHE: const	CES277	AVERAGE HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFARM PAYROLLS - CONSTRUCTION
131	AHE: mfg	CES278	AVERAGE HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFARM PAYROLLS - MANUFACTURING
132	Consumer expect	HHSNTN	U. OF MICH. INDEX OF CONSUMER EXPECTATIONS(BCD-83)

Acknowledgements

The views expressed are solely the responsibility of the individual authors and should not be interpreted as reflecting the views of the Executive Board of Sveriges Riksbank. The paper has benefited from comments received following presentations of previous versions at the Far Eastern Meeting of the Econometric Society, Beijing, China, July 2006, the 13th International Conference on Computing in Economics and Finance, Montréal, Canada, June 2007, the Research Forum: New Developments in Dynamic Factor Modelling, Centre for Central Banking Studies of the Bank of England, London, October 2007, the Far Eastern Meeting of the Econometric Society, Singapore, July 2008, the Conference on Factor Structures for Panel and Multivariate Time Series Data, Maastricht, September 2008, and seminars at De Nederlandsche Bank, the Sveriges Riksbank and the University of Groningen. Helpful comments were also received from three referees.

References

[**1**] Bai, J. (2003), "Inferential theory for factor models of large dimensions", *Econometrica*, **71**, 135–171.

[**2**] Bai, J. and S. Ng (2002), "Determining the number of factors in approximate factor models", *Econometrica*, **70**, 191–221.

[**3**] Bai, J. and S. Ng (2007), "Determining the number of primitive shocks in factor models", *Journal of Business and Economic Statistics*, **25**, 52–60.

[**4**] Bai, J. and S. Ng (2008a), "Large dimensional factor analysis", *Foundations and Trends in Econometrics*, **3**, 89–163.

[**5**] Bai, J. and S. Ng (2008b), "Forecasting economic time series using targeted predictors", *Journal of Econometrics*, **146**, 304–317.

[**6**] Boivin, J. and S. Ng (2006), "Are more data always better for factor analysis?", *Journal of Econometrics*, **132**, 169–194.

[**7**] Burnham, K.P. and D.R. Anderson (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edition, Springer, New York.

[**8**] Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000), "The generalized dynamic-factor model: Identification and estimation", *The Review of Economics and Statistics*, **82**, 540–554.

[**9**] Forni, M. and M. Lippi (2001), "The generalized factor model: Representation theory", *Econometric Theory*, **17**, 1113–1141.

[**10**] Goodwin, G.C. and R.L. Payne (1977), *Dynamic System Identification: Experiment Design and Data Analysis*, Academic Press, New York, London.

[**11**] Greene, W.H. (2008), *Econometric Analysis*, sixth edition, Pearson / Prentice Hall, Upper Saddle River, NJ.

[**12**] Hallin, M. and R. Liška (2007), "Determining the number of factors in the general dynamic factor model", *Journal of the American Statistical Association*, **102**, 603–617.

[**13**] Inklaar, R.C., J.P.A.M. Jacobs, and W.E. Romp (2004), "Business cycle indexes: Does a heap of data help?", *Journal of Business Cycle Measurement and Analysis*, **1**, 309–336.

[**14**] Jacobs, J.P.A.M. and P.W. Otter (2008), "Determining the number of factors and lag order in dynamic factor models: A minimum entropy approach", *Econometric Reviews*, **27**, 385–397.

[**15**] Kullback, S. and R.A. Leibler (1951), "On information and sufficiency", *Annals of Mathematical Statistics*, **22**, 79–96.

[**16**] Muirhead, R.J. (1982), *Aspects of Multivariate Statistical Theory*, Wiley, Mew York.

[**17**] Onatski, A. (2009), "Testing hypotheses about the number of factors in large factor models", *Econometrica*, **77**, 1447–1479.

[**18**] Otter, P.W. and J.P.A.M. Jacobs (2006), "On information in static and dynamic factor models", Working Paper #2006/5, CCSO [presented at

the Far Eastern Meeting of the Econometric Society, July 2006, Beijing, China].

[**19**] Otter, P.W., J.P.A.M. Jacobs, and A.H.J. den Reijer (2011), "A criterion for the number of factors in a data-rich environment", Mimeo, University of Groningen.

[**20**] Stock, J.H. and M.W. Watson (2002), "Forecasting using principal components from a large number of predictors", *Journal of the American Statistical Association*, **97**, 1167–1179.

[**21**] Stock, J.H. and M.W. Watson (2005), "Implications of dynamic factor models for VAR analysis", Working paper 11467, National Bureau of Economic Research.

[**22**] Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society. Series B*, **58**, 267–288.

[**23**] Young, T.Y. and T.W. Calvert (1974), *Classification, Estimation and Pattern Recognition*, Elsevier, New York; London.

[**24**] Zou, H. and T. Hastie (2005), "Regularization and variable selection via the elastic net", *Journal of Royal Statistical Society, Series B*, **67**, 301–320.